2700-101
VMD:glj

SPEECH RECOGNITION FOR RECOGNIZING SPEAKER-INDEPENDENT, CONTINUOUS SPEECH

BACKGROUND OF THE INVENTION

1. Field of the Invention

[0001] The present invention relates generally to speech recognition, and more particularly to real-time speech recognition for recognizing speaker-independent, connected or continuous speech.

2. Description of the Background Art

[0002] Speech recognition refers to the ability of a machine or device to receive, analyze and recognize human speech. Speech recognition is also often referred to as voice recognition. Speech recognition may potentially allow humans to interface with machines and devices in an easy, quick, productive and reliable manner. Accurate and reliable speech recognition is therefore highly sought after.

[0003] Speech recognition gives humans the capability of verbally generating documents, recording or transcribing speech, and audibly controlling devices. Speech recognition is desirable because speech occurs at a much faster rate than manual operations, such as typing on a keyboard or operating controls. A good typist can type about 80 words per minute, while typical speech can be in the range of about 200 or more words per minute.

[0004] In addition, speech recognition can allow remote control of electronic devices. Many applications exist for impaired persons who cannot operate conventional devices, such as persons who are at least partially paralyzed, blind, or medicated. For example, a computer or computer operated appliances could be speech controlled.

[0005] Moreover, speech recognition may be used for hands-free operation of conventional devices. For example, one current application is the use of speech recognition for operating a cellular phone, such as in a vehicle. This may be desirable because the driver's attention should stay on the road.

[0006] Speech recognition processes and speech recognition devices currently exist. However, there are several difficulties that have prevented speech recognition from becoming practical and widely available. The main obstacle has been the wide variations in speech between persons. Different speakers have different speech characteristics, making speech recognition difficult or at best not satisfactorily reliable. For example, useful speech recognition must be able to identify not only words but also small word variations. Speech recognition must be able to differentiate between homonyms by using context. Speech recognition must be able to recognize silence, such as gaps between words. This may be difficult if the speaker is speaking rapidly and running words together. Speech recognition systems

may have difficulty adjusting to changes in the pace of speech, changes in speech volume, and may be frustrated by accents or brogues that affect the speech.

[0007] Speech recognition technology has existed for some time in the prior art, and has become fairly reasonable in price. However, it has not yet achieved satisfactory reliability and is not therefore widely used. For example, as previously mentioned, devices and methods currently exist that capture and convert the speech into text, but generally require extensive training and make too many mistakes.

[0008] FIG. 1 shows a representative audio signal in a time domain. The audio signal is generated by capture and conversion of an audio stream into an electronic voice stream signal, usually through a microphone or other sound transducer. Generally, audible sound exists in the range of about 20 hertz (cycles) to about 20 kilohertz (kHz). Speech is a smaller subset of frequencies. The electronic voice stream signal may be filtered and amplified and is generally digitized for processing.

[0009] FIG. 2 shows the voice stream after it has been converted from the time domain into the frequency domain. Conversion to the frequency domain offers advantages over the time domain. Human speech is generated by the mouth and the throat, and contains many different harmonics (it is generally not composed of a single component frequency). The audible

speech signal of FIG. 2, therefore, is composed of many different frequency components at different amplitude levels. In the frequency domain, the speech recognition device may be able to more easily analyze the voice stream and detect meaning based on the frequency components of the voice stream.

[0010] FIG. 3 shows how the digitized frequency domain response may be digitally represented and stored. Each digital level may represent a frequency or band of frequencies. For example, if the input voice stream is in the range of 1 kilohertz (kHz) to 10 kHz, and is separated into 128 frequency spectrum bands, each band (and corresponding frequency bin) would contain a digital value or amplitude for about 70 Hz of the speech frequency spectrum. This value may be varied in order to accommodate different portions of the audible sound spectrum. Speech does not typically employ all of the frequencies in the audible frequency range of 20 Hz to 20 kHz. Therefore, a speech recognition device may analyze only the frequencies from 1 kHz to 10 kHz, for example.

[0011] Once the voice stream has been converted to the frequency domain, an iterative statistical look-up may be performed to determine the parts of speech in a vocalization. The parts are called phonemes, the smallest unit of sound in any particular language. Various languages use phonemes that are not utilized in any other language. The English language designates

about 34 different phonemes. The iterative statistical look-up employed by the prior art usually uses hidden Markov modeling (HMM) to statistically compare and determine the phonemes. The iterative statistical look-up compares multiple portions of the voice stream to stored phonemes in order to try to find a match. This generally requires multiple comparisons between a digitized sample and a phoneme database and a high computational workload. Therefore, by finding these phonemes, the speech recognition device can create a digital voice stream representation that represents the original vocalizations in a digital, machine-usable form.

[0012] The main difficulty encountered in the prior art is that different speakers speak at different rates and therefore the phonemes may be stretched out or compressed. There is no standard length phoneme that a speech recognition device can look for. Therefore, during the comparison process, these time-scale differences must be compensated for.

[0013] In the prior art, the time-scale differences are compensated for by using one of two approaches. In the dynamic time warping process, a statistical modeling stretches or compresses the wave form in order to find a best fit match of a digitized voice stream segment to a set of stored spectral patterns or templates. The dynamic time warping process uses a

procedure that dynamically alters the time dimension to minimize the accumulated distance score for each template.

[0014] In a second prior art approach, the hidden Markov model (HMM) method characterizes speech as a plurality of statistical chains. The HMM method creates a statistical, finite-state Markov chain for each vocabulary word while it trains the data. The HMM method then computes the probability of generating the state sequence for each vocabulary word. The word with the highest accumulated probability is selected as the correct identification. Under The HMM method, time alignment is obtained indirectly through the sequence of states.

[0015] The prior art speech recognition approaches have drawbacks. One drawback is that the prior art approach is not sufficiently accurate due to the many variations between speakers. The prior art speech recognition suffers from mistakes and may produce an output that does not quite match.

[0016] Another drawback is that the prior art method is computationally intensive. Both the dynamic time warping approach and the HMM statistical approach require many comparisons in order to find a match and many iterations in order to temporally stretch or compress the digitized voice stream sample to fit samples in the phoneme database.

[0017] There have been many attempts in the prior art to increase speech recognition accuracy and/or to decrease

computational time. One way of somewhat reducing computational requirements and increasing accuracy is to limit the library of phonemes and/or words to a small set and ignore all utterances not in the library. This is acceptable for applications requiring only limited speech recognition capability, such as operating a phone where only a limited number of vocal commands are needed. However, it is not acceptable for general uses that require a large vocabulary (i.e., normal conversational speech).

[0018] Another prior art approach is a speaker-dependent speech recognition wherein the speech recognition device is trained to a particular person's voice. Therefore, only the particular speaker is recognized, and that speaker must go through a training or "enrollment" process of reading or inputting a particular speech into the speech recognition device. A higher accuracy is achieved without increased cost or increased computational time. The drawback is that use of speaker-dependent voice recognition is limited to one person, requires lengthy training periods, may require a lot of computation cycles, and is limited to only applications where the speaker's identity is known apriori.

[0019] What is needed, therefore, are improvements in speech recognition technology.

SUMMARY OF THE INVENTION

[0020] A speech recognition device is provided according to one embodiment of the invention. The speech recognition device comprises an I/O device for accepting a voice stream and a frequency domain converter communicating with the I/O device. The frequency domain converter converts the voice stream from a time domain to a frequency domain and generates a plurality of frequency domain outputs. The speech recognition device further comprises a frequency domain output storage communicating with the frequency domain converter. The frequency domain output storage comprises at least two frequency spectrum frame storages for storing at least a current frequency spectrum frame and a previous frequency spectrum frame. A frequency spectrum frame storage of the at least two frequency spectrum frame storages comprises a plurality of frequency bins storing the plurality of frequency domain outputs. The speech recognition device further comprises a processor communicating with the plurality of frequency bins and a memory communicating with the processor. A frequency spectrum difference storage in the memory stores one or more frequency spectrum differences calculated as a difference between the current frequency spectrum frame and the previous frequency spectrum frame. At least one feature storage is included in the memory for storing at least one feature extracted from the voice stream. At least one transneme table is included in the memory, with the at least one transneme table including a plurality of transneme table entries and with a transneme table

entry of the plurality of transneme table entries mapping a predetermined frequency spectrum difference to at least one predetermined transneme of a predetermined verbal language. At least one mappings storage is included in the memory, with the at least one mappings storage storing one or more found transnemes. At least one transneme-to-vocabulary database is included in the memory, with the at least one transneme-to-vocabulary database mapping a set of one or more found transnemes to at least one speech unit of the predetermined verbal language. At least one voice stream representation storage is included in the memory, with the at least one voice stream representation storage storing a voice stream representation created from the one or more found transnemes. The speech recognition device calculates a frequency spectrum difference between a current frequency spectrum frame and a previous frequency spectrum frame, maps the frequency spectrum difference to a transneme table, and converts the frequency spectrum difference to a transneme if the frequency spectrum difference is greater than a predetermined difference threshold. The speech recognition device creates a digital voice stream representation of the voice stream from one or more transnemes thus produced.

[0021] A method for performing speech recognition on a voice stream is provided according to a first method embodiment of the invention. The method comprises the steps of determining one or more candidate transnemes in the voice stream, mapping the one or more candidate transnemes to a transneme table to convert the one

or more candidate transnemes to one or more found transnemes, and mapping the one or more found transnemes to a transneme-to-vocabulary database to convert the one or more found transnemes to one or more speech units.

[0022] A method for performing speech recognition on a voice stream is provided according to a second method embodiment of the invention. The method comprises the step of calculating a frequency spectrum difference between a current frequency spectrum frame and a previous frequency spectrum frame. The current frequency spectrum frame and the previous frequency spectrum frame are in a frequency domain and are separated by a predetermined time interval. The method further comprises the step of mapping the frequency spectrum difference to a transneme table to convert the frequency spectrum difference to at least one transneme if the frequency spectrum difference is greater than a predetermined difference threshold. A digital voice stream representation of the voice stream is created from one or more transnemes thus produced.

[0023] A method for performing speech recognition on a voice stream is provided according to a third method embodiment of the invention. The method comprises the step of performing a frequency domain transformation on the voice stream upon a predetermined time interval to create a current frequency spectrum frame. The method further comprises the step of normalizing the current frequency spectrum frame. The method further comprises the step of calculating a frequency spectrum

difference between the current frequency spectrum frame and a previous frequency spectrum frame. The method further comprises the step of mapping the frequency spectrum difference to a transneme table to convert the frequency spectrum difference to at least one found transneme if the frequency spectrum difference is greater than a predetermined difference threshold. The method therefore creates a digital voice stream representation of the voice stream from one or more found transnemes thus produced.

[0024] The above and other features and advantages of the present invention will be further understood from the following description of the preferred embodiments thereof, taken in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0025] FIG. 1 shows a representative audio signal in a time domain;

[0026] FIG. 2 shows the voice stream after it has been converted from the time domain into the frequency domain;

[0027] FIG. 3 shows how the digitized frequency domain response may be digitally represented and stored;

[0028] FIG. 4 shows a speech recognition device according to one embodiment of the invention;

[0029] FIG. 5 shows detail of a frequency domain output storage;

[0030] FIG. 6 is a flowchart of a first speech recognition method embodiment according to the invention;

[0031] FIG. 7 shows a frequency spectrum frame according to a first embodiment of a frequency domain conversion;

[0032] FIG. 8 shows a frequency spectrum frame according to a second embodiment of the frequency domain conversion;

[0033] FIG. 9 is a flowchart of a second speech recognition method embodiment;

[0034] FIG. 10 shows a first frequency spectrum frame obtained at a first point in time;

[0035] FIG. 11 shows a second frequency spectrum frame obtained at a second point in time;

[0036] FIG. 12 shows how the frequency domain transformation may be processed using overlapping frequency domain conversion windows;

[0037] FIG. 13 is a flowchart of a third speech recognition method embodiment;

[0038] FIGS. 14-16 show a frequency normalization operation on a current frequency spectrum frame; and

[0039] FIGS. 17-18 show an amplitude normalization operation on a current frequency spectrum frame.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0040] FIG. 4 shows a speech recognition device 400 according to one embodiment of the invention. The speech recognition device 400 includes an input/output (I/O) device 401, a frequency domain converter 406, frequency domain output storage 410, a processor 414, and a memory 420.

[0041] The speech recognition device 400 of the invention performs speech recognition and converts a voice stream input into a digital voice stream representation. The voice stream representation comprises a series of symbols that digitally represents the voice stream and may be used to recreate the voice stream.

[0042] The speech recognition is accomplished by finding transnemes in the voice stream and converting the found transnemes into speech units of a predetermined verbal language. A speech unit may be a word, a portion of a word, a phrase, an expression, or any other type of verbal utterance that has an understood meaning in the predetermined verbal language.

[0043] A phoneme is generally described as being the smallest unit of sound in any particular language. Each vocalized phoneme is a distinct sound and therefore may be characterized by a substantially unique frequency domain response, substantially over the duration of the vocalization of the phoneme. In the English language, it is generally accepted that there are about 34 phonemes that are used to create all parts of the spoken

language. There are less than 100 identified phonemes in all languages combined.

[0044] A transneme is a transition between the phoneme (or allophone) components of human speech. There are approximately 10,000 transnemes. Transnemes are therefore smaller subunits or components of speech, and are used by the invention to produce a speech recognition that is speaker-independent and that operates on connected speech (i.e., the speaker can talk normally and does not have to take care to voice each word separately and distinctly).

sub a2)

[0045] The speech recognition of the invention does not attempt to find and identify phonemes, as is done in the prior art. Instead, the speech recognition device 400 searches for transitions between and within phonemes (i.e., transnemes), with such transitions generally being shorter in duration than phonemes. Moreover, because a transneme is defined by two temporally adjacent phonemes or parts of phonemes, the number of transnemes is approximately equal to the square of the number of phonemes (i.e., 100x100=10,000). Identification of the transneme components of speech therefore achieved a greater efficiency, accuracy, and resolution than the various speech recognition techniques of the prior art.

[0046] The I/O device 401 may be any type of device that is capable of accepting an audio signal input that includes a voice

stream. The I/O device 401 provides the audio signal input to the speech recognition device 400. The I/O device 401 may accept a digitized voice stream or may include a digitizer, such as an analog to digital converter (not shown) if the incoming audio signal is in analog form. The I/O device 401 may accept a voice stream that is already compressed or that has already been converted into the frequency domain.

[0047] The I/O device 401 may be any type of input device, including a microphone or sound transducer, or some other form of interface device. The I/O device 401 may additionally be a radio frequency receiver or transceiver, such as a radio frequency front-end, including an antenna, amplifiers, filters, down converters, etc., that produce an audio signal. This may include any type of radio receiver, such as a cell phone, satellite phone, pager, one or two-way radios, etc. Furthermore, the I/O device 401 may be a TV receiver, an infrared receiver, an ultrasonic receiver, etc. Alternatively, the I/O device 401 may be an interface, such as an interface to a digital computer network, such as the Internet, a local area network (LAN), etc., may be an interface to an analog network, such as an analog telephone network, etc.

[0048] In addition to accepting a voice stream, the I/O device 401 may be capable of outputting a voice stream representation. For example, the speech recognition may be used to compress the

voice stream for transmission to another device. For example, the speech recognition of the invention may be used in a cell phone. The voice stream (or other sound input) may be received by the I/O device 401 and may be converted into text or speech representation symbols by the speech recognition device 400. The speech representation symbols may then be transmitted to a remote location or device by the I/O device 401. At the receiving device, the speech representation symbols may be converted back into an audio output. This use of the invention for audio compression greatly reduces bandwidth requirements of the speech transmission.

[0049] The frequency domain converter 406 communicates with the I/O device 401 and converts the incoming voice stream signal into a frequency domain signal (See FIGS. 2 and 3). The frequency domain converter 406 may be, for example, a discrete Fourier transform device or a fast Fourier transform device (FFT) that performs a Fourier transform on the time domain voice stream. Alternatively, the frequency domain converter 406 may be a filter bank employing a plurality of filters, such as band pass filters that provide a frequency spectrum output, or may be a predictive coder.

[0050] The frequency domain converter 406 generates a plurality of outputs, with each output representing a predetermined frequency or frequency band. The plurality of

outputs of the frequency domain conversion are referred to herein as a frequency spectrum frame, with a frequency spectrum frame comprising a plurality of amplitude values that represent the frequency components of the voice stream over the predetermined frequency conversion window period (see FIG. 10, for example, and also see the discussion accompanying FIG. 13). The number of outputs may be chosen according to a desired frequency band size and according to a range of audible frequencies desired to be analyzed. In one embodiment, the frequency domain converter 406 generates 128 outputs for a predetermined frequency domain conversion window.

[0051] FIG. 5 shows detail of the frequency domain output storage 410. The frequency domain output storage 410 communicates with the frequency domain converter 406. The frequency domain output storage 410 is a memory that comprises at least two frequency spectrum frames 410a and 410b. Alternatively, the frequency domain output storage 410 may be eliminated and the plurality of outputs from the frequency domain converter 406 may be stored in the memory 420.

[0052] Each frequency spectrum frame 410a, 410b, etc., stores a set of digital values V_1-V_N that represents the amplitudes or quantities of frequency band components present in the voice stream over the predetermined frequency conversion window. Each frame contains N bins, with N corresponding to the number of

frequency domain transformation outputs generated by the frequency domain converter 406. Each bin therefore contains a frequency domain conversion output value V. The sets of digital values V_1-V_N in successive frequency spectrum frames, such as those in frames 410a and 410b, may be analyzed in order to process the input voice stream.

[0053] The processor 414 may be any type of processor, and communicates with the frequency domain output storage 410 in order to receive the frequency spectrum frame or frames contained within the frequency domain output storage 410. The processor 414 is also connected to the memory 420 and is optionally connected to the I/O device 401 whereby the processor may receive a digitized time domain signal. The processor 414 may be connected to the I/O device 401 in order to extract features from the time domain signal, such as pitch and volume features. These speech features may be used for adding punctuation, emphasis, etc., to the voice stream representation output, for example, and may also be used to normalize the frequency spectrum frames (the normalization is discussed below in the text accompanying FIGS. 14-18).

[0054] The memory 420 communicates with the processor 414 and may be any type of storage device, including types of random access memory (RAM), types of read-only memory (ROM), magnetic tape or disc, bubble memory, optical memory, etc. The memory 420

may be used to store an operating program that includes a speech recognition algorithm according to various aspects of the invention. In addition, the memory 420 may include variables and data used to process the speech and may hold temporary values during processing. The memory 420, therefore, may include a spectrum difference storage 421, a feature storage 422, at least one transneme table 423, at least one mappings storage 424, at least one transneme-to-vocabulary database 427, and at least one voice stream representation 428. The memory 420 may optionally include a frequency spectrum frame storage (not shown) for storing one or more frequency spectrum frames, such as the output of the frequency domain converter 406.

[0055] The spectrum difference storage 421 stores at least one frequency spectrum difference calculated from current and previous frequency spectrum frames. The frequency spectrum difference, therefore, is a set of values that represents a change in spectral properties of the voice stream.

[0056] The feature storage 422 contains at least one feature extracted from the voice stream, such as a volume or pitch feature, for example. The feature may have been obtained from the voice stream in either the time domain or the frequency domain. Data stored in the feature storage 422 may be used to normalize a frequency spectrum frame and to aid in grammar interpretation such as by adding punctuation. In addition, the

features may be used to provide context when matching a found transneme to speech units, words, or phrases.

[0057] The at least one transneme table 423 contains a plurality of transnemes. Transnemes are used to analyze the input voice stream and are used to determine the parts of speech therein in order to create a voice stream representation in digital form. For example, the phrase, "Hello, World" is made up of 11 transnemes ("[]" represents silence). If smaller frame sizes are utilized, transnemes can be smaller parts of phonemes, with multiple transnemes per phoneme.

[] - H

H - E

E - L

L - OW

OW - []

[] - W

W - EH

EH - R

R - L

L - D

D - []

[0058] Transnemes are dictated by the sounds produced by the human vocal apparatus. Therefore, transnemes are essentially independent of the verbal language of the speaker. However, a

particular language may only contain a subset of all of the existing transnemes. As a result, the transneme table 423 may contain only the transnemes necessary for a predetermined verbal language. This may be done in order to conserve memory space. In applications where language translation or speech recognition of multiple verbal languages is required, the transneme table may likely contain the entire set of transnemes.

[0059] The mappings storage 424 stores one or more mappings produced by comparing one or more frequency spectrum differences to at least one transneme table 423. In other words, when transnemes are found through the matching process, they are accumulated in the mappings storage 424.

[0060] The transneme-to-vocabulary database 427 maps found transnemes to one or more speech units, with a speech unit being a word, a portion of a word, a phrase, or any utterance that has a defined meaning. By using the transneme-to-vocabulary database 427, the speech recognition device 400 may compare groupings of one or more found transnemes to the transneme-to-vocabulary database 427 in order to find speech units and create words and phrases.

[0061] The transneme-to-vocabulary database 427 may contain entries that map transnemes to one or more verbal languages, and may therefore convert transnemes into speech units of one or more predetermined verbal languages. In addition, the speech

recognition device 400 may include multiple transneme-to-vocabulary databases 427, with additional transneme-to-vocabulary databases 427 capable of being added to give additional speech recognition capability in other languages.

[0062] The voice stream representation storage 428 is used to store found words and phrases as part of the creation of a voice stream representation. For example, the voice stream representation 428 may accumulate a series of symbols, such as text, that have been constructed from the voice stream input.

[0063] The speech recognition device 400 is speaker-independent and reliably processes and converts connected speech (connected speech is verbalized without any concern or effort to separate the words or talk in any specific manner in order to aid in speech recognition). The speech recognition device 400 therefore operates on connected speech in that it can discern breaks between words and/or phrases and without excessive computational requirements. A reduced computational workload may translate to simpler and less expensive hardware.

[0064] Another difference between the invention and the prior art is that gaps or silences between phonemes are detected and removed by the frequency spectrum differences. If a spectrum difference evaluates to be approximately zero, that means that the spectrum frame has not changed appreciably since the last frequency domain transformation. Therefore, a transition from

one phoneme to another (i.e., a transneme) has not occurred and as a result the particular time sample may be ignored.

[0065] The use of frequency spectrum frame differencing eliminates the need for the complex and inefficient dynamic time warping procedure of the prior art, and therefore generally requires only about one cycle per comparison of a spectrum difference to a database or table. Therefore, unlike the prior art dynamic time warping and statistical modeling, the speech recognition device 400 of the present invention does not need to perform a large amount of comparisons in order to account for time variations in the voice stream.

[0066] The speech recognition device 400 may be a specialized device constructed for speech recognition or may be integrated into another electronic device. The speech recognition device 400 may be integrated into any manner of electronic device, including cell phones, satellite phones, conventional land-line telephones (digital and analog), radios (one and two-way), pagers, personal digital assistants (PDAs), laptop and desktop computers, mainframes, digital network appliances and workstations, etc. Alternatively, the speech recognition device 400 may be integrated into specialized controllers or general purpose computers by implementing software to perform speech recognition according to the present invention. Speech recognition and control may therefore be added to personal

computers, automobiles and other vehicles, factory equipment and automation, robotics, security devices, etc.

[0067] FIG. 6 is a flowchart 600 of a first speech recognition method embodiment according to the invention. In step 606, the method determines one or more candidate transnemes in a voice stream. This is accomplished by analyzing the voice stream in a frequency domain and comparing frequency spectrum frames (captured over predetermined time periods) in order to determine one or more candidate transnemes. The frequency spectrum frames may be obtained for overlapping time periods (windows).

[0068] As previously described, a transneme is a transition between phonemes or allophones, and a candidate transneme may be determined by finding a significant spectrum variation in the frequency domain. Therefore, a comparison in the frequency domain may be performed between a current frequency spectrum frame (containing frequency components of predetermined frequencies or frequency bands) to a previous frequency spectrum frame in order to determine voice stream frequency changes over time. Periods of silence or periods of substantially no frequency change are ignored.

[0069] The I/O device 401 passes a digitized voice stream to the frequency domain converter 406 and to the processor 414. The processor 414 may extract predetermined speech features from the digitized time domain voice stream. These predetermined speech

features may be used to add punctuation and may also be used to normalize the frequency spectrum frames based on a detected volume and pitch of the speaker. For example, a tonality rise by the speaker may signify a question or query, signifying a question mark or other appropriate punctuation in the completed voice stream representation.

[0070] The frequency domain converter 406 converts the digitized voice stream into a plurality of frequency domain signal outputs and stores them in the frequency domain output storage 410. The frequency domain converter 406 is preferably a Fourier transform device that performs Fourier transforms on the digitized input voice stream. The outputs are preferably in the form of an array of values representing the different frequency bands present in the voice stream. Alternatively, the frequency domain converter 406 may be any other device that is capable of converting the time domain signal into the frequency domain, such as a filter bank comprising a plurality of filters, such as band pass filters, for example.

[0071] FIG. 7 shows a frequency spectrum frame 700 according to a first embodiment of the frequency domain conversion. In the first embodiment, the frequency spectrum frame 700 comprises a plurality of contiguous frequency bands that substantially covers a predetermined portion of the audible frequency spectrum.

[0072] FIG. 8 shows a frequency spectrum frame 800 according to a second embodiment of the frequency domain conversion. In the second embodiment, the frequency spectrum frame 800 comprises a plurality of substantially individual frequencies or a plurality of non-contiguous frequency bands. Although some of the frequencies in the predetermined portion of the audible frequency spectrum are ignored, the frame 800 may still adequately reflect and characterize the various frequency components of the voice stream. By using only portions of the predetermined portion of the audible frequency spectrum and not using contiguous frequency bands, the amount of data processed in comparing frequency spectrum characteristics and finding transnemes may be further reduced. The result is a decrease in computational processing requirements and storage requirements.

[0073] As a further part of determining a transneme, the processor 414 accesses the frequency domain outputs in the frequency domain output storage 410 and creates a frequency spectrum difference. When a frequency spectrum difference evaluates to be non-zero, a transition between phonemes has occurred and the frequency spectrum difference has been used to determine a candidate transneme in the input voice stream.

[0074] In step 612, a candidate transneme is mapped to the transneme table 423 (or other data conversion device) in order to determine whether it is a valid transneme. By using one or more

transneme tables 423, a candidate transneme (i.e., a valid frequency spectrum difference) is converted into a found transneme.

[0075] In step 615, the found transneme is mapped to at least one transneme-to-vocabulary database 427 (or other data conversion device) and are converted to one or more speech units. The speech units may comprise words, portions of words, phrases, or any other utterance that has a recognized meaning.

[0076] The method 600 therefore converts the voice stream (or other audio input) into a digital voice stream representation. The digital voice stream representation produced by the speech recognition comprises a series of digital symbols, such as text, for example, that may be used to represent the voice stream. The voice stream representation may be stored or may be used or processed in some manner.

[0077] Speech recognition according to the invention has many uses. The speech recognition of the invention converts a voice stream input into a series of digital symbols that may be used for speech-to-text conversion, to generate commands and inputs for voice control, etc. This may encompass a broad range of applications, such as dictation, transcription, messaging, etc. The speech recognition of the invention may also encompass voice control, and may be incorporated into any type of electronic device or electronically controlled device. Furthermore, the

speech recognition of the invention may analyze any type of non-speech audio input and convert it to a digital representation if an equivalent set of transneme definitions is known. For example, the speech recognition method could find possible applications in music.

[0078] The speech recognition of the invention may additionally be used to perform a highly effective compression on the voice stream. This is accomplished by converting the voice stream into a voice stream representation comprising a series of symbols. Due to the highly efficient and relatively simple conversion of the voice stream into digital symbols (such as numerical codes, including, for example, ASCII symbols for English letters and punctuation), the speech recognition of the invention may provide a highly effective audio signal compression.

[0079] Digitally captured and transmitted speech typically contains only frequencies in the 4 kHz to 10 kHz range and requires a data transmission rate of about 12 kbits per second. However, employing the speech recognition of the invention, speech may be transmitted at a data rate of about 120 bits per second to about 60 bits per second. The result is a data compression of about 100 to 1 to about 200 to 1. This allows a device to decrease its data rate and hardware requirements while

simultaneously allowing greater transmission quality (i.e., by capturing more of the 20 Hz to 20 kHz audible sound spectrum).

[0080] The voice stream compression may be advantageously used in a variety of ways. One application may be to compress a voice stream for transmission. This may include wireless transmission of the voice stream representation, such as a radio transmission, an infrared (IR) or optical transmission, an ultrasonic transmission, etc. The voice stream compression may therefore be highly useful in communication devices such as cellular phones, satellite phones, pagers, radios, etc. Alternatively, the transmission may be performed over some form of transmission media, such as wire, cable, optical fiber, etc. Therefore, the voice stream representation may be used to more efficiently communicate data over conventional analog telephone networks, digital telephone networks, digital packet networks, computer networks, etc.

[0081] Another compression application may be the use of speech recognition to compress data for manipulation and storage. A voice stream may be converted into a series of digital symbols in order to drastically reduce storage space requirements. This may find advantageous application in areas such as answering machines, voice messaging, voice control of devices, etc.

[0082] In yet another application, by including an appropriate transneme-to-vocabulary table or tables 427, the speech

recognition may perform a language translation function. The speech recognition device 400 may receive a voice stream of a first verbal language and, through use of an appropriate transneme-to-vocabulary table or tables 427, may convert that voice stream into a voice stream representation of a second language.

[0083] FIG. 9 is a flowchart 900 of a second speech recognition method embodiment according to the invention. In step 903, a frequency spectrum difference is calculated. The frequency spectrum difference is calculated between two frequency spectrum frames in order to determine whether a transneme has occurred. The voice stream must have already been processed in some manner in order to create a frequency domain signal or representation. This may also include pre-processing such as amplification, filtering, and digitization.

[0084] FIG. 10 shows a first frequency spectrum frame 1000 obtained at a first point in time T_1. FIG. 11 shows a second frequency spectrum frame 1100 obtained at a second point in time T_2. From these two frames, it can be seen that the frequency components of the voice stream have changed over the time period T_2-T_1(see dashed lines). Therefore, the frequency spectrum difference will reflect these spectral changes and may represent a transneme.

[0085] As part of the frequency spectrum difference calculation, the difference may be compared to a predetermined difference threshold to see if a transneme has occurred. If the frequency spectrum difference is less than or equal to the predetermined difference threshold, the difference may be judged to be essentially zero and the current frequency spectrum frame may be ignored. A next frequency spectrum frame may then be obtained and processed.

[0086] The predetermined difference threshold takes into account noise effects and imposes a requirement that the frames must change by at least the predetermined difference threshold in order to be a valid transneme. In one embodiment, the predetermined difference threshold is about 5% of average amplitude of base frequency spectrum bin over a less than 100 millisecond frame size, although the predetermined difference threshold may range from about 3% to about 7%.

[0087] A frequency spectrum difference is preferably calculated about every 10 milliseconds due to the average time duration of a phoneme, but may be varied according to conditions or a desired resolution. For example, a frequency spectrum difference may be calculated more often in order to increase the resolution of the transneme differentiation and to potentially increase accuracy. However, the computational workload will increase as a consequence.

[0088] FIG. 12 shows how the frequency domain transformation may be processed using overlapping frequency domain conversion windows F1, F2, F3, etc. By using overlapping windows, the method ensures that no voice stream data is missed. On the contrary, each data point in the voice stream is preferably processed twice due to the overlap. In addition, the overlap ensures that the analysis of each transneme is more accurate by comparing each voice stream feature to the transneme table 423 more than once.

[0089] In step 906, the frequency spectrum difference is mapped to some form of reference in order to determine the transneme that has occurred. The reference will typically be one or more transneme tables or databases, with a predetermined frequency difference mapping to a predetermined transneme. The predetermined frequency spectrum difference may map to more than one predetermined transneme. In use, there may be up to 5 or 10 transnemes in the transneme table 423 that may substantially match the predetermined frequency spectrum difference. An optional size optimization function can be performed to compress that for memory-sensitive applications (at a small expense of processing cycles). A final found transneme may be determined through use of the feature data and by comparison of transneme groupings to the transneme-to-vocabulary database 427. The found

transneme or transnemes may be stored and accumulated in the mappings storage 424.

[0090] Preferably, the mapping of the frequency spectrum differences may be done using a hash-like vector-distance metric, which finds the best fit difference-to-transneme equivalence. The transneme table 423 may be constructed from experimentally derived transneme data that comprises manually evaluated speech waveforms that are broken down into transnemes. Alternatively, the transneme table 423 may be created by converting existing phoneme data into transnemes.

[0091] In step 908, the found transnemes are used to create a voice stream representation in digital form. The found transnemes are preferably processed as groupings, such as a grouping of 10-20 transnemes, for example. After a grouping of transnemes has been accumulated, a free-text-search-like lookup is preferably performed against the transneme-to-vocabulary database 427, using an inverted-index technique to find the best-fit mappings of found transnemes to a word or phrase. Many duplications of words may exist in the transneme-to-vocabulary database 427 in order to accommodate various groupings and usages of words, homonym-decoding, speaker-restarts, etc.

[0092] The inverted index is an index into a plurality of text entries, such as a database, for example, that is produced during a text search of the database. The inverted index search result indicates the database entry where the search term may be found

and also indicates a location within the text of the matched entry.

[0093] For example, consider a simple database (shown in lower case for simplicity) containing four text entries:

```
I love you
god is love
love is blind
blind justice
```

[0094] As an example, a search is performed for each of the words in the database. Using an inverted index search to index the search results by (entry, offset within the entry), the search result indexes returned from the database search might appear as:

```
blind    (3,8);(4,0)
god      (2,0)
I        (1,0)
is       (2,4);(3,5)
justice  (4,6)
love     (1,2);(2,7);(3,0)
you      (1,7)
```

[0095] The word "blind" is therefore in database entry 3 ("love is blind"), starting at character 8, and is also in entry 4 ("blind justice"), starting at the first character.

[0096] To find documents containing both "is" and "love," the search result indexes are examined to find intersections between the entries. In this case, both entries 2 and 3 contain the two words, so the database entries 2 and 3 both contain the two

search terms. Therefore, each entry number of the "is" result may be compared to the "love" result, with an intersection occurring when a database entry number is present in both indexes. In an additional capability, an inverted index search can quickly find documents where the words are physically close by comparing the character offsets of a result having at least one intersecting entry.

[0097] The inverted index search technique may be applied to the vocabulary lookup of the transnemes by adding an additional transneme field to each database entry of the transneme-to-vocabulary database 427. The transneme-to-vocabulary database 427 might then have the form:

Fld1 Doc#	Fld2 Clear Text:	Fld3 Transneme Version
1	"I love you"	00Ah AhEE EE00 00LL LLUh UhVv VvYY YYoo ooww ww00
2	"god is love"	00Gg GgAh Ahdd ddih ihZZ ZZLL LLUh UhVv Vv00
3	"love is blind"	00LL LLUh UhVv Vvih ihZZ ZZ00 00Bb BbLl LlAh AhEE EENn NnDd Dd00
4	"blind justice"	00Bb BbLl LlAh AhEE EENn NnDd Dd00 00Dj DjUh UhSs Sstt ttih ihss ss00

Where

Field 1: Record # (== row number, == document number)
Field 2: Clear text (in the predetermined verbal language)
Field 3: Transneme version of text
Field 4: Sub-index of each word-to-transneme mapping (optional)

[0098] Once candidate transnemes are identified from frequency spectrum differences, they may be used as inverted index query arguments and may be queried against the Field3 transneme

version. Therefore, a transneme code or representation obtained from the transneme table 423 may be compared to the transneme versions in the transneme-to-vocabulary database 427 until a match is found. Any database entries which match the query are returned, along with a relative relevance ranking for each result.

[0099] The particular word that matched the search term may also be identified in the result field. The word match may be determined from a separate "child" table, or may be determined from an additional field of the same table. For example, the database entry:

```
1     "I love you"     00Ah AhEE EE00 00LL LLUh UhVv VvYY YYoo ooww ww00
```

may also be indexed for each transneme-to-word mapping. The transneme-to-vocabulary database 427 might therefore also contain the entries:

```
5     "I"         00Ah AhEE EE00
6     "love"      00LL LLUh UhVv
7     "you"       YYoo ooww ww00
```

This secondary mapping may be as efficiently done as a post-processing scan of the returned transnemes to identify the word boundaries. The efficiency of the secondary mapping is linked to the number of words in the returned clear text phrases.

[0100] After the current grouping of found transnemes has been identified, successive additional iterations may be performed to

continue the speech recognition process. At the start of each iteration, the frequency domain conversion window may be advanced. The frequency domain conversion window may be advanced by only a portion of its length, so that a current frequency domain conversion overlaps a previous conversion. In a preferred embodiment, the frequency conversion window is about 10 milliseconds in duration and it is advanced about half a window length. An overlapping lookup is performed against the transneme-to-vocabulary database 427. This may be done in order to ensure that no voice stream data is overlooked and to increase reliability of the analysis through multiple comparisons. The overlapping conversions may also be used to clarify a match, using a context, and may be used to vote for best matches. The overlapping frequency domain conversions may be used to create a listing for a transneme grouping of the search lookups of the transneme-to-vocabulary database 427, with the transneme matches contained in the listing being organized and ranked according to relevance.

[0101] FIG. 13 is a flowchart 1300 of a third speech recognition method embodiment according to the invention. In step 1301, predetermined features are extracted from the voice stream. The predetermined features may include the volume or amplitude of the voice stream, the pitch (frequency), etc. This feature extraction may be done on a time domain version of the voice stream, before the input voice stream is converted into the

frequency domain. Alternatively, the feature extraction may be mathematically extracted from the input voice stream in the frequency domain. For example, in a cellular phone application, a data stream received by the cellular phone may already be in the frequency domain, and the feature extraction may be mathematically performed on the received digital data in order to calculate the volume and the pitch of the speaker.

[0102] In step 1303, a frequency domain transformation is performed on the digitized voice stream input to produce the frequency domain output. The output comprises a plurality of frequency band values that represent various frequency components within the voice stream.

[0103] The frequency domain transformation may be performed by a FFT device. Alternatively, the frequency domain transformation may be performed by a filter bank comprising a plurality of filters, such as band pass filters. The output of the filter bank is a plurality of frequency band outputs having amplitudes that represent the frequency components within each band.

[0104] The frequency domain conversion may be performed for a predetermined time period or window, capturing frequency domain characteristics of the voice stream for a window of time. The frequency domain conversion window is preferably about 10 milliseconds in size. However, this size may be varied, with the size being chosen to accommodate factors such as a desired

resolution of the speech. In addition, the language of the speaker may be a factor in choosing the conversion window size. The size of the conversion window must be chosen so as to balance the desired resolution against computational cycles, hardware complexity, etc., because an increase in speech recognition resolution may be accompanied by an increase in computational cycles and/or hardware/memory size.

[0105] In step 1305, the frequency domain transformation output is stored. This is preferably done in a plurality of frequency bins, with each bin corresponding to a particular predetermined frequency or frequency band.

[0106] In step 1309, the frequency domain outputs are normalized. This may include a frequency normalization and an amplitude normalization, for example. The normalization is performed using speech features that have been extracted from the time domain digitized voice signal. The voice features may include pitch, volume, speech rate, etc. The normalization, therefore, adjusts the values in the frequency bins in order to maintain frequency spectrum frames that are essentially constant (except for frequency components of the voice stream). This accommodates variations in volume and frequency. For example, even if the speaker is speaking very loudly, the overall volume should not matter and should not affect the speech recognition

result. Likewise, a change in frequency/pitch by the speaker should not affect the speech recognition.

[0107] FIGS. 14-16 show a frequency normalization operation on a current frequency spectrum frame. In order to prevent changes in tonality (frequency) from affecting the transneme determination, a base frequency may be used for a frequency normalization. The base frequency is the frequency band (or bin) containing the greatest amplitude (i.e., the largest frequency component of the speech at that time). During frequency normalization, the contents of the frequency bins may be shifted up or down in order to maintain the base frequency in a substantially constant frequency bin location.

[0108] FIG. 14 shows a previous frequency spectrum frame 1400 and a frequency bin set containing the corresponding values. FIG. 15 shows a current frequency spectrum frame 1500 and a frequency bin set containing the corresponding values. It can be seen that the base frequency of the current frequency spectrum frame 1500 is higher in frequency (tonality) than the base frequency of the previous frequency spectrum frame 1400. This may be due to various factors, such as emotion or emphasis on the part of the speaker, etc.

[0109] FIG. 16 shows a frequency normalization comprising a frequency shift of the current frequency spectrum frame 1500, forming a new current frame 1600. This may be done merely by

shifting the values (V_1-V_N, for example) in the frequency bin set. This may entail dropping one or more values in order to accommodate the shift. A predetermined frequency shift threshold may be used to prevent excessive frequency normalization, allowing only a limited amount of shifting. In addition to normalizing the current frame, the normalization value may be saved and may be used in a word lookup to determine context and punctuation.

[0110] FIGS. 17 and 18 show an amplitude normalization operation on a current frequency spectrum frame. The figures show two successive frequency spectrum frames 1700 and 1800, where the only difference between the frames is an amplitude difference "a". This amplitude difference may be due to a change in volume of the speaker. However, this change in volume could potentially be seen as a transition between phonemes. Therefore, a normalization is desirable so that the speech recognition is essentially independent of the volume of the speaker. The normalization may be accomplished by uniformly increasing or decreasing the values in all of the frequency bins in order to substantially match the amplitude of the previous frequency spectrum frame. In this example, the value "a" is subtracted from all of the frequency spectrum frame values V_1-V_N in all of the frequency bins.

[0111] In step 1316, a frequency spectrum difference is calculated. As previously discussed, the frequency spectrum difference is calculated between two frequency spectrum frames in order to determine whether a transneme has occurred. The frequency spectrum difference therefore is a set of values that show the difference in frequency components from the current frequency spectrum frame and a previous frequency spectrum frame.

[0112] In step 1320, the frequency spectrum difference is mapped to the transneme table 423 (or other reference or database) in order to determine a found transneme.

[0113] In step 1325, the transnemes are used to create a voice stream representation in digital form stored within the voice recognition device or computer. The voice stream representation output is a data stream composed of text or digital representations, such as a series of symbols, complete with some suggested punctuation.

[0114] While the invention has been described in detail above, the invention is not intended to be limited to the specific embodiments as described. It is evident that those skilled in the art may now make numerous uses and modifications of and departures from the specific embodiments described herein without departing from the inventive concepts.